Assessment of Contaminant Fate in Heterogeneous Porous Media using Importance Sampling Monte Carlo Methods

Zhiming Lu,[1] Andrew V. Wolfsberg,[1] and Chuan Lu[2]

Zhiming Lu, Computational Earth Sciences Group (EES-16), Los Alamos National Laboratory, Los Alamos, NM 87545, USA. (zhiming@lanl.gov)

[1] Computational Earth Sciences Group (EES-16), MS T003, Los Alamos National Laboratory, Los Alamos, NM 87545

[2] Energy and Geoscience Institute, The University of Utah, Salt Lake City, UT 84108, USA

Abstract. In many applications, one may be interested in estimating the probability of a released solute reaching a particular target region in a randomly heterogeneous porous medium. The conventional Monte Carlo method can be used for such a purpose. However, if the probability to be estimated is relatively small, an extremely large number of realizations (or particles) may be required to obtain a reasonably accurate estimate. In this study, we introduce a new approach, importance sampling Monte Carlo simulation (ISMC), to efficiently estimate such a small probability. In the conventional Monte Carlo simulations (CMC), the probability of interest is derived from an ensemble of all possible trajectories that are attributed to heterogeneity of the hydraulic conductivity field (assuming the local dispersion is negligible). In the importance sampling approach, such trajectories are taken from a modified ensemble so that more solute particles will reach the target region. We do so by adding an artificial spatially-varying velocity field to the true velocity field. Since the samples are taken from a biased ensemble, the outputs from simulations are then weighted in such a way that the bias introduced by sampling from the modified ensemble will be exactly corrected. The general procedure of this importance sampling approach as well as its applicability to subsurface transport problems has been illustrated using several examples for which analytical solutions are available. The comparison of results from analytical solutions, the conventional Monte Carlo method, and the importance sampling approach demonstrates that the latter is compu-

tational much more efficient than the conventional Monte Carlo simulations,

especially when the probability of interest in very small.

1. Introduction

Geological formations are ubiquitously heterogeneous and thus fluid flow and solute transport processes in these formations are normally treated as stochastic processes. Stochastic approaches to flow and transport in heterogeneous porous media have been extensively studied in the last two decades, and many theoretical models have been developed [*Dagan*,1989; *Gelhar*, 1993; *Zhang*, 2002]. These stochastic models include, for instance, Monte Carlo simulations, moment-equation methods [*Zhang*, 1998; *Guadagnini and Neuman*, 1999; *Tartakovsky et al.*, 1999], spectral methods [*Yeh and Gelhar*, 1985; *Li and McLaughlin*, 1991], KL-based moment methods [*Zhang and Lu*, 2004; *Lu and Zhang*, 2004a; *Lu and Zhang*, 2004b], *Lu and Zhang*, 2007], and probability collocation methods [*Li and Zhang*, 2007]. Except for some special cases in which analytical solutions are available [*Osnes*, 1995; *Lu and Zhang*, 2005], most of these models rely on numerical solutions of the stochastic partial differential equations. A detailed comparison on performance of the Monte Carlo method and moment-equation methods can be found in [*Lu and Zhang*, 2004b].

The moment-equation-based approach in many cases works well for relatively large variations of medium properties [*Tartakovsky et al.*, 1999]. However, in general it is restricted to small variabilities of medium properties. More importantly, in the moment method, one has to solve covariance equations as many times as the number of nodes, although it may render accurate solutions with coarser numerical grids (of fewer nodes) than the grid in the Monte Carlo method, Thus, this approach can also be computationally demanding, in particular, for simulating field-scale problems. The moment-equation method based on

Karhunen-Loeve decomposition (KLME) does not require to solve the covariance equations directly and only needs to solve a few hundred modes of the state variables (e.g., hydraulic head or solute concentration) on relatively coarse numerical grids [*Zhang and Lu*, 2004]. The KLME method has been extended to solve non-linear unsaturated flow problems [*Yang et al.*, 2004], conditional simulations [*Lu and Zhang*, 2004a], non-stationary fields with zonation [*Lu and Zhang*, 2007], and multi-phase flow problems [*Chen et al.*, 2006]. This method may be used to conduct three-dimensional field-scale simulations [*Lu and Zhang*, 2006] if the heterogeneous random field (say, hydraulic conductivity) can be easily decomposed, for example, if the covariance of the random field is a separable exponential function for which the field can be analytically decomposed. The computational cost of the Karhunen-Loeve decomposition at the field scale can be high if the covariance function has to be decomposed numerically.

The Monte Carlo method is still widely used in practical applications. This approach is conceptually straightforward, which is accomplished by generating a large number of *equally likely* random realizations of the parameter fields, solving deterministic flow and/or transport equations for each realization, and averaging the results over all realizations to obtain statistical moments of state variables. This approach has an advantage of applying to a very broad range of both linear and nonlinear flow and transport problems, but has a number of potential drawbacks [*Tartakovsky et al.*, 1999].

A major disadvantage of the Monte Carlo method, among others, is the requirement for large computational effort. This is partial due to fine numerical grids and small time steps required to properly resolve high frequency space-time fluctuations in random parameters, which leads to large computational effort for each individual realization. This

is particularly true if both physical and chemical heterogeneities, as well as uncertainties in initial and boundary conditions, are considered. In addition, to ensure sample moments of state variables converge to their theoretical ensemble values, a large number of Monte Carlo simulations are often required (typically a few thousand realizations, depending on the degree of medium heterogeneity). This is because the estimation error of the Monte Carlo method is inversely proportional to the square root of the sample size. The estimation error decreases with the sample size but the rate of convergence is rather slow.

The importance sampling method is one of effective techniques that can reduce the estimation error much fast. This technique has been extensively used in communication theory in simulating rare events [*Chen et al.*, 1993; *Sadowsky*, 1993]. Applications in other fields include, for example, computing probabilities of low-energy nuclear collisions [*Mazonka et al.*, 1998] and calculating rate constants for transitions between stable states [*Dellago et al.*, 1998]. A detailed mathematical framework for importance sampling can be found in *Glynn* [1989] and a thorough review on this topic can be found in *Smith* [1997].

Lu and Zhang [2003a] applied the importance sampling method to solve flow and transport problems in random porous media. They illustrated the method using two one-dimensional flow and transport problems by assuming that the input random variable is a random constant. By the importance sampling method, they were able to substantially reduce computational effort required by the CMC method. However, the method has been limited to cases with random variables rather than correlated random fields.

In this paper, we are interested in estimating the probability of a released solute reaching a particular target region in a randomly heterogeneous porous medium. The conventional Monte Carlo method can be used for such a purpose, but the computational cost could

be very high if the probability to be estimated is relatively small. In the conventional Monte Carlo simulations, the probability of interest is derived from an ensemble of all possible trajectories that are attributed to heterogeneities of the hydraulic conductivity field (assuming local dispersion is negligible). In the importance sampling approach, such trajectories are taken from a modified ensemble so that more solute particles will reach the target region. Because the samples are taken from a biased ensemble, the outputs from simulations are then weighted in such a way that the bias introduced by sampling from the modified ensemble will be exactly corrected. The comparison of results from analytical solutions, the CMC method, and ISMC approach demonstrates that the latter is computationally much more efficient than the conventional Monte Carlo method, especially when the probability of interest in very small.

This paper is organized as follows. We first introduce some basic concepts of the importance sampling method in Section 2. The Lagrangian equation for solute transport in heterogeneous porous media is revisited in Section 3. In Section 4, statistical distributions of Lagrangian trajectories are derived and the importance sampling technique is applied to the problem. Two numerical examples are given in Section 5, followed by a short discussion and conclusion in Section 6.

2. Basic Concepts of Importance Sampling Techniques

The purpose of applying the importance sampling technique is to obtain an accurate estimate of a quantity of interest with fewer samples than required by the CMC method. Two major steps are involved in the importance sampling technique. First, one needs to modify the process being simulated such that it is much easy to simulate the modified process, which is characterized by a new probability density function other than the origi-

nal density function. In modeling, instead of taking samples from the original probability density function, samples are taken from this new probability density function, called importance density function, such that some "important" regions of the sample space get more samples. The effectiveness of this approach depends on the choice of the importance density function. Thus a fundamental issue in implementing the importance sampling technique is how to select a good density function. Second, because samples are taken from a biased density function, one needs to correct the bias by averaging the output from different samples (realizations) using weights that are related to both the original and modified density functions, such that the mean of the quantity being estimated is preserved.

Let Ω be a probability space and $f_T(\mathbf{X})$ be a probability density function defined on this space, where subscript T stands for the target distribution. We want to derive the moments of function $\theta(\mathbf{X})$, where θ is a specified, deterministic function (called score function). For example, if $\mathbf{X}$ is a particle's trajectory or a path between its release location and a pre-defined target region, $\theta(\mathbf{X})$ can be a function with two values, being one if the particle reaches the target region or zero otherwise. By definition, the mean of θ can be written as

$$\langle\theta\rangle_T = E_T[\theta(\mathbf{X})] = \int_\Omega \theta(\mathbf{x}) f_T(\mathbf{x}) d\mathbf{x}, \tag{1}$$

where E_T represents statistical expectation under the target density function f_T, and $\langle\theta\rangle_T$ denotes the mean of θ under density function f_T. To estimate the mean from (1) using the CMC method, one randomly samples a sequence of $\mathbf{X}_i$, $i = \overline{1, N}$, from the density

function $f_T(\mathbf{x})$ and computes the sample mean

$$\overline{\theta} = \frac{1}{N}\sum_{i=1}^{N}\theta(\mathbf{X}_i). \tag{2}$$

Since $\mathbf{X}_i$, $i = \overline{1,N}$, are independent identically distributed (i.i.d.) random variables, it can be shown that $E_T[\overline{\theta}] = \langle\theta\rangle_T$, i.e., $\overline{\theta}$ is an unbiased estimator of $\langle\theta\rangle_T$. The variance of θ can be computed as

$$\sigma_\theta^2 = E\{[\theta - \langle\theta\rangle_T]^2\} = \int_\Omega \theta^2(\mathbf{x}) f_T(\mathbf{x}) d\mathbf{x} - \langle\theta\rangle_T^2. \tag{3}$$

Since the score function satisfies $\theta^2 = \theta$, the first term in (3) is the same as the mean and thus the variance can be written as $\sigma_\theta^2 = \langle\theta\rangle_T(1 - \langle\theta\rangle_T)$. For a given level of desired accuracy ϵ, the number of required Monte Carlo simulations, $N_{\epsilon,MC}$, can be estimated from the following equation [*Lu and Zhang*, 2003a]:

$$N_{\epsilon,MC} = \frac{\sigma_\theta^2}{\epsilon^2\langle\theta\rangle_T^2} = \frac{1 - \langle\theta\rangle_T}{\epsilon^2\langle\theta\rangle_T}. \tag{4}$$

This equation indicates that the number of required Monte Carlo simulations is inversely proportional to the mean value $\langle\theta\rangle_T$, which means that, if $\langle\theta\rangle_T$ is very small, a large number of Monte Carlo simulations will be required. Because the statistical error of the mean estimation in (2) is inversely proportional to $\sqrt{N}$, if we want to reduce the error by a factor of two we have to increase the sample size by a factor of four. Certainly, the convergence rate of the CMC method is rather slow. The importance sampling technique is a way that reduces the estimation variance and thus reduces the statistical error much faster than the CMC method does.

Suppose we sample $\mathbf{X}_i$, $i = \overline{1,N}$, from another (sampling) density function $f_S(\mathbf{x})$ rather than the original target density function $f_T(\mathbf{x})$, where $f_S(\mathbf{x})$ is zero only if $f_T(\mathbf{x})$ is zero. To preserve the mean (i.e., to correct the bias), we define a weight function $w(\mathbf{x}) =$

$f_T(\mathbf{x})/f_S(\mathbf{x})$, and the expectation under density function $f_S(\mathbf{x})$ then can be determined

$$\begin{aligned}\langle\theta\rangle_S = E_S[w(\mathbf{X})\theta(\mathbf{X})] &= \int_\Omega \theta(\mathbf{x})w(\mathbf{x})f_S(\mathbf{x})d\mathbf{x} \\ &= \int_\Omega \theta(\mathbf{x})f_T(\mathbf{x})d\mathbf{x} \\ &= \langle\theta\rangle_T, \end{aligned} \tag{5}$$

which means that the mean remains the same though the samples are taken biasly from the density function $f_S(\mathbf{x})$.

From (5) one can construct an estimator based on samples $\mathbf{X}_i$, $i = \overline{1,N}$:

$$\overline{\theta} = \frac{1}{N}\sum_{i=1}^{N}\theta(\mathbf{X}_i)w(\mathbf{X}_i), \tag{6}$$

i.e., the contribution of sample $\mathbf{X}_i$ in $\overline{\theta}$ is weighted by $w(\mathbf{x}_i)$ and the bias due to sampling from the biased importance density function has been corrected. The variance of the estimator can be written as

$$\begin{aligned}\sigma^2_{\theta,S} = E_S[\theta^2w^2] - (E_S[w\theta])^2 &= \int_\Omega \theta^2w^2f_S(\mathbf{x})d\mathbf{x} - \langle\theta\rangle_S^2 \\ &= \int_\Omega \theta^2wf_T(\mathbf{x})d\mathbf{x} - \langle\theta\rangle_S^2, \end{aligned} \tag{7}$$

where E_S denotes the expectation under the sampling density function f_S. By noticing that the second terms in (3) and (7) are the same, it is seen that the estimate variance can be reduced by choosing the weighting function w such that on averaging it is much smaller than unity.

For a given accuracy ϵ, the number of required simulations can be derived similarly as $N_{\epsilon,ISMC} = \sigma^2_{\theta,S}/\epsilon^2\langle\theta\rangle_S^2$. The ratio of variances $\gamma = N_{\epsilon,MC}/N_{\epsilon,ISMC}$ is a measure of the efficiency (performance) of the importance sampling method, which depends on the choice of the importance density function. This means that, comparing to the CMC, sampling from the importance density function may allow us to estimate the mean with a small

sample size for a given accuracy, or with a much better accuracy for a given sample size. In fact, as will be shown later, the Monte Carlo approach based on importance sampling techniques makes it possible to solve some problems that cannot be efficiently solved by the conventional Monte Carlo simulations.

3. Lagrangian Transport Equations

In this study, we are interested in probabilities that a particle released at a known position $\mathbf{X}_0$ reaches some particular areas. For a particle of conservative species originating from location $\mathbf{X}_0$ at $t = t_0$, its trajectory is described by the following kinetic equation:

$$\frac{d\mathbf{X}(t)}{dt} = \mathbf{V}[\mathbf{X}(t)], \tag{8}$$

with the initial condition $\mathbf{X}(0) = \mathbf{X}_0$, where $\mathbf{X}(t)$ stands for particle's position at time t and $\mathbf{V}[\mathbf{X}(t)]$ denotes the (Lagrangian) velocity of the particle. It should be emphasized that even if in the case that the flow field is steady state, the particle's (Lagrangian) velocity $\mathbf{V}[\mathbf{X}(t)]$ may still be time-dependent if the (Eulerian) flow field is spatially non-stationary, which may be caused by, for example, nonstationarity of soil properties or appropriate boundary conditions.

Following *Lu and Zhang* [2003b], when the Eulerian velocity is treated as a random space function, the particle's Lagrangian velocity and its position should also be treated as random functions. Let us denote $\mathbf{X}(t)$ as $\mathbf{X}$ when there is no confusion. We may decompose the particle position $\mathbf{X}$ into its mean and fluctuation: $\mathbf{X} = \langle \mathbf{X} \rangle + \mathbf{X}'$. With Taylor expansion we may expand the Lagrangian velocity $\mathbf{V}(\mathbf{X})$ at the mean position $\langle \mathbf{X} \rangle$ as

$$\mathbf{V}(\mathbf{X}) = \mathbf{V}(\langle \mathbf{X} \rangle) + (\mathbf{X}' \cdot \nabla)\mathbf{V}(\langle \mathbf{X} \rangle) + \frac{1}{2}(\mathbf{X}' \cdot \nabla)^2 \mathbf{V}(\langle \mathbf{X} \rangle) + \cdots. \tag{9}$$

Substituting (9) into (8) leads to

$$\frac{d\mathbf{X}}{dt} = \langle \mathbf{V}(\langle \mathbf{X} \rangle) \rangle + \mathbf{V}'(\langle \mathbf{X} \rangle) + (\mathbf{X}' \cdot \nabla)\langle \mathbf{V}(\langle \mathbf{X} \rangle) \rangle, \tag{10}$$

with the initial condition $\mathbf{X}(0) = \mathbf{X}_0$, or

$$\frac{d\mathbf{X}}{dt} = \langle \mathbf{V}(\langle \mathbf{X} \rangle) \rangle + \mathbf{V}'(\langle \mathbf{X} \rangle) + \mathbf{B}(t)\mathbf{X}', \tag{11}$$

where $B_{ij}(t) = \partial \langle V_i(\mathbf{X}_t) \rangle / \partial X_{t,j}|_{\langle \mathbf{X} \rangle}$ is the derivative of particle's mean velocity with respect to its trajectory, evaluated at the current position $\langle \mathbf{X}(t) \rangle$. *Lu and Zhang* [2003b] considered the general case of solute spreading in a spatially nonstationary flow field, where the mean flow may vary spatially in both magnitude and direction, and derived the second moments of $\mathbf{X}$ on the basis of the general expression (11). The special case of $B_{ij}(t) \equiv 0$ represents the uniform mean flow condition [e.g., *Dagan*, 1989]. For simplicity in our discussion, we ignore the last term in (11), which is a good approximation if $B_{ij}(t)$ is small [*Indelman and Rubin*, 1996; *Sun and Zhang*, 2000; *Lu and Zhang*, 2003b].

4. Statistical Properties of Trajectories

Consider the solute transport process that is described by the Lagrangian transport equation as derived in the previous section:

$$\frac{d\mathbf{X}}{dt} = \langle \mathbf{V}(\langle \mathbf{X}(t) \rangle) \rangle + \mathbf{V}'(\langle \mathbf{X}(t) \rangle), \tag{12}$$

for $0 \le t \le \tau$. Any single realization of this process can be described by a trajectory $\mathbf{X}(t)$, $0 \le t \le \tau$, with $\mathbf{X}(0) = \mathbf{X}_0$. We denote all possible trajectories by an ensemble T

$$T(\tau) = \left\{ \mathbf{X}(t) \,\middle|\, \dot{\mathbf{X}} = \langle \mathbf{V} \rangle + \mathbf{V}',\ \mathbf{X}(0) = \mathbf{X}_0; 0 \le t \le \tau \right\}. \tag{13}$$

We are interested in the probability of a particle released at source $\mathbf{X}_0$ reaching a specific region Ω_0, a subdomain of the simulation domain D, at time τ (Fig. 1). Now we introduce

a functional $\Theta[\mathbf{X}(t)]$ which equal to 1 if $\mathbf{X}(\tau) \in \Omega_0$ and 0 otherwise. The probability being sought, $P = \langle \Theta \rangle_T$, is the average of this functional over the ensemble T defined in (13). We can compute P by sampling randomly from ensemble T and counting the number of hits:

$$P \approx \frac{1}{N} \sum_{n=1}^{N} \Theta[\mathbf{X}_n^T(t)], \tag{14}$$

where $\mathbf{X}_n^T$ is the n^{th} independent simulation (sampling) from ensemble T. For a very small value of P, a very large number of simulations are required.

Now we consider a modified version of (12)

$$\frac{d\mathbf{X}}{dt} = \langle \mathbf{V}(\langle \mathbf{X}(t) \rangle) \rangle + \Delta\mathbf{V} + \mathbf{V}'(\langle \mathbf{X}(t) \rangle), \tag{15}$$

where $\Delta\mathbf{V}$ is a velocity field added to the original velocity field. Later we will see that this added velocity field will affect the convergence rate of the importance sampling method. The ensemble of trajectories corresponding to (15) can be written as:

$$S(\tau) = \left\{ \mathbf{X}(t) \,\middle|\, \dot{\mathbf{X}} = \langle \mathbf{V} \rangle + \Delta\mathbf{V} + \mathbf{V}',\ \mathbf{X}(0) = \mathbf{X}_0; 0 \le t \le \tau \right\}. \tag{16}$$

If we define the weight function as

$$w(\mathbf{X}(t)) = \frac{p_T(\mathbf{X}(t))}{p_S(\mathbf{X}(t))}, \tag{17}$$

where $p_T(\mathbf{X}(t))$ and $p_S(\mathbf{X}(t))$ are the probabilities of sampling $\mathbf{X}(t)$ from ensembles T and S, respectively, the probability to be sought can be written as

$$P = \langle w\Theta \rangle_S \approx \frac{1}{N} \sum_{n=1}^{N} w(\mathbf{X}_n^S(t))\, \Theta(\mathbf{X}_n^S(t)). \tag{18}$$

Here $\mathbf{X}_n^S(t)$ is the n^{th} independent simulation using the modified version of the transport equation. The estimation variance can be computed similarly using

$$\sigma^2_{ISMC} = \langle w^2\Theta^2 \rangle - \langle w\Theta \rangle_S^2 \approx \frac{1}{N} \sum_{n=1}^{N} w^2(\mathbf{X}_n^S(t))\, \Theta(\mathbf{X}_n^S(t)) - P^2. \tag{19}$$

Now the problem becomes how to evaluate the weight function w, or more precisely, how to estimate probabilities p_T and p_S from equations (12) and (15), which differ by an added velocity field. It should be noted that one does not need to actually sample trajectories $\mathbf{X}(t)$ from ensemble T but only need to evaluate probabilities $p_T(\mathbf{X})$ and $p_S(\mathbf{X})$ for samples $\mathbf{X}$ taken from ensemble S.

More general, we consider the following equation

$$\frac{d\mathbf{X}}{dt} = \mathbf{U} + \mathbf{V}'(\langle \mathbf{X}(t) \rangle), \tag{20}$$

where U is the mean velocity. Equations (12) and (15) are two instances of (20). Discretize the trajectory $\mathbf{X}(t)$, $0 \le t \le \tau$, into M equal increments in time as $\mathbf{X}_m = \mathbf{X}(m\delta t)$, $m = \overline{1, M}$, $\delta t = \tau/M$ (see Fig. 1). The probability that the trajectory starts from $\mathbf{X}_0$ at $t = 0$ and passes $(\mathbf{X}_1, \cdots, \mathbf{X}_M)$ at $\delta t, 2\delta t, \cdots, \tau$ can be found similarly as in *Mazonka et al.* [1998]:

$$P(\mathbf{Y}|\mathbf{X}_0) = \left[(2\pi\delta t)^d \det(\mathbf{D})\right]^{-M/2} \exp[-A(\mathbf{Z})], \tag{21}$$

where d is the dimension of the problem, $\mathbf{D}$ is velocity covariance, $\mathbf{Z} = (\mathbf{X}_0, \mathbf{Y}) = (\mathbf{X}_0, \mathbf{X}_1, \cdots, \mathbf{X}_M)$, and

$$A(\mathbf{Z}) = \frac{1}{2\delta t} \sum_{m=0}^{M-1} (\mathbf{X}_{m+1} - \mathbf{X}_m - \mathbf{U}\delta t)^T D^{-1} (\mathbf{X}_{m+1} - \mathbf{X}_m - \mathbf{U}\delta t). \tag{22}$$

Substituting U in (21)-(22) by V and $V + \Delta V$ respectively, we can find $p_T(\mathbf{X}(t))$ and $p_S(\mathbf{X}(t))$, and thus the weight function as

$$w(\mathbf{X}(t)) = \frac{p_T(\mathbf{X}(t))}{p_S(\mathbf{X}(t))} = \exp(-\Delta A), \tag{23}$$

where ΔA, in the limit of $\delta t \to 0$, can be computed as

$$\Delta A = \frac{1}{2} \int_0^\tau (2\mathbf{V}' + \Delta\mathbf{V}) \mathbf{D}^{-1} \Delta\mathbf{V} dt \tag{24}$$

which can be computed by solving

$$\frac{d\mathbf{F}}{dt} = \frac{1}{2}\left(2\mathbf{V}' + \Delta\mathbf{V}\right)\mathbf{D}^{-1}\Delta\mathbf{V} \tag{25}$$

with initial condition $\mathbf{F}(0) = 0$, and then setting $\Delta A = \mathbf{F}(\tau)$. It should be noted that, the velocity perturbation $\mathbf{V}'$ is spatially dependent in simulating any single realization and it also varies from realization to realization. For each realization (trajectory) sampled from ensemble $S(\tau)$, all quantities in the integrand of (24), i.e., $\mathbf{V}'$, $\Delta\mathbf{V}$ and D, are evaluated along this trajectory, and thus ΔA, or equivalently weight w, can be different from realization to realization.

5. Illustrative Examples

5.1. Travel time statistics in one-dimensional problems

In the first example, we revisit the one-dimensional transport problem presented in *Lu and Zhang* [2003a]. One of the major reasons for their choosing such a simple example is that an analytical solution for this problem is available, and therefore it can serve as the basis for comparing the efficiency of the CMC and ISMC methods.

In their example, they considered one-dimensional saturated flow with prescribed deterministic influx q on the left end and deterministic constant head on the right. The hydraulic conductivity was also a deterministic constant. It was assumed that porosity ϕ was a random constant following a normal distribution $N(\mu_\phi, \sigma_\phi^2)$ and they were interested in the probability of particle's travel time less than a given value T_0. Other parameters in the example were given as: $q = 0.1\ m/day$ and $L = 100.0\ m$. For this particular problem, the value of conductivity in the domain and the constant head value at the right end were not relevant.

Since porosity ϕ is a random variable, so is the travel time $t = L\phi/q$. In fact, the travel time is also normally distributed: $t \sim N(L\mu_\phi/q, L^2\sigma_\phi^2/q^2)$. The probability $P(t < T_0)$ can be found explicitly as

$$P_{T_0} = P(t < T_0) = \Phi\left(\frac{T_0 q/L - \mu_\phi}{\sigma_\phi}\right), \tag{26}$$

where $\Phi(x)$ is the cumulative density function (*cdf*) of the standard normal distribution. For any given T_0, the probability computed from (26) will be used to evaluate the accuracy and computational efficiency of the CMC and ISMC methods.

In the CMC approach, one samples a sequence of ϕ_i, $i = \overline{1, N}$. If ϕ_i is less than or equal to qT_0/L, the score function $\theta(\phi_i)$ is set to unity. Otherwise, $\theta(\phi_i)$ is set to zero. The scores from all N samples are collected and the sample mean is calculated: $\overline{P}_{T_0,MC} \approx (1/N)\sum_{i=1}^{N}\theta(\phi_i)$. If one wants to estimate P_T using this approach, for small T_0, a large number of MC simulations are required to obtain a reasonably accurate results. For example, if one wants to estimate the probability that a particle's travel time is less than 100 days, i.e., $P(t < 100)$, which is 1.31×10^{-11}, one needs to conduct 7.64×10^{12} simulations to retain an accuracy of 10% [*Lu and Zhang*, 2003a].

Lu and Zhang [2003a] used an importance sample method based on Mean Translation (MT) to shift the mean of the porosity to a lower value that is related to T_0. For example, for $T_0 = 100$ days, the mean of the porosity was shifted to $\mu_\phi = 0.1$. Certainly, it is much easy to get sample values around $\phi = 0.1$ (or equivalently $T = 100$ days) from $N(\mu_\phi = 0.1, \sigma_\phi^2)$ than from $N(\mu_\phi = 0.3, \sigma_\phi^2)$. After taking samples ϕ_i, $i = \overline{1, N}$ from the importance density function, the probability to be sought can be found:

$$\overline{P}_{T,IS} \approx \frac{1}{N}\sum_{i=1}^{N}\theta(\phi_i)w(\phi_i) = \frac{1}{N}\sum_{i=1}^{N}\frac{f(\phi_i)}{f_1(\phi_i)}\theta(\phi_i), \tag{27}$$

where f and f_1 represent the probability density function of the original and modified (shifted) normal distributions, respectively. By using the importance sampling method, they were able to estimate the probability within $\pm 10\%$ accuracy with less then 1,000 simulations.

Numerical results from the CMC and ISMC methods are tabulated in Table 1, where the number in parentheses is the number of simulations used to compute P_T. As expected, for an accuracy of 10%, 1,000 simulations is enough for the ISMC approach for T_0 value as low as 50 days, while a significantly large number of runs is required for the CMC approach to achieve the same level of accuracy even for $T_0 = 200$ days. For $T_0 <= 150$ days, an accurate estimation of P_T using the CMC method could not be obtained.

Now we apply the importance sampling techniques presented in this paper to the same problem by first defining two ensembles $T(t)$ and $S(t)$. Since the velocity is a random constant over the one-dimensional domain, an ensemble of trajectories can be equivalently considered as an ensemble of travel time:

$$T(t) = \left\{ t = \frac{L\phi}{q} \,\middle|\, \phi \sim N(\mu_\phi, \sigma_\phi^2) \right\}, \tag{28}$$

and

$$S(t) = \left\{ t = \frac{L}{q/\phi + \Delta v} \,\middle|\, \phi \sim N(\mu_\phi, \sigma_\phi^2) \right\}. \tag{29}$$

Now we need to find the weight function w by evaluating two probability density functions (PDF) of the travel time under the original (target) and modified distributions. The PDF of the original distribution can be found by taking the derivative of (26):

$$f_T(T_0) = \frac{q}{L\sigma_\phi} f_n(z); \quad z = \frac{qT_0/L - \mu_\phi}{\sigma_\phi}, \tag{30}$$

where f_n is the PDF of the standard normal distribution. Similarly, by adding a velocity component to the original random velocity q/ϕ, the travel time can be written as $t = L/(q/\phi + \Delta v)$, and the probability that travel time is less than T_0 can be written as

$$P_S = P(t < T_0) = P\left(\frac{L}{q/\phi + \Delta v} < T_0\right) = \Phi(u), \tag{31}$$

where $u = [qT_0/(L - T_0\Delta v) - \mu_\phi]/\sigma_\phi$. The corresponding density function can be derived from (31) by taking its derivative with respect to T_0:

$$f_S(T_0) = \frac{qL}{(L - T_0\Delta v)^2\sigma_\phi} f_n(u). \tag{32}$$

From the expressions of f_T and f_S, one may find the optimized Δv by minimizing f_T/f_S, which yields

$$\Delta v = \frac{L}{T_0} - \frac{2q}{\sqrt{\mu_\phi^2 + 8\sigma_\phi^2} + \mu_\phi}. \tag{33}$$

If one is interested in cases with $T_0 > q/\mu_\phi$, a similar expression can be derived. This equation indicates that, for given μ_ϕ and σ_ϕ^2, the optimized Δv is a function of T_0. The physical meaning of this expression can be easily seen by neglecting σ_ϕ^2 (for small variance as in our case), which leads to $\Delta v = L/T_0 - q/\mu_\phi$. It is seen that the second term is the first-order mean velocity, while the first term is the "required velocity" such that the mean travel time will be T_0. Therefore, Δv in (33) is indeed an extra velocity component needed. The values of $\Delta v \approx L/T_0 - q/\mu_\phi$ for different T_0 are listed in Table 1. The calculated probability values using the importance sampling method are listed in Table 1 as P_{IS} with the number of simulations in parentheses. It is seen from the table that, using the ISMC approach, we are able to estimate the probabilities using 1000 samples only while retaining a high accuracy, even for the probability value as low as in the order of 10^{-17}. It should be pointed out that the computed probability values in Table 1 may

be slightly different from run to run, depending on the selected random number generator and initial random seeds.

5.2. Two-dimensional problems with random velocity fields

In the second example, we consider a two-dimensional domain $\Omega = [0, 1000] \times [-300, 300]$ (in any consistent units), which is discretized into a grid of 50×30 with an element size of 20×20. Although the methodology can be applied to any velocity field, for simplicity, in this example we choose a stationary mean velocity $\langle \mathbf{v} \rangle = (v_1, v_2)^T = (1, 0)^T$, and a velocity fluctuation of $\mathbf{v}' = (c_1\xi_1, c_2\xi_2)^T$, where c_1 and c_2 are two constants and ξ_1 and ξ_2 are standard Gaussian random variables. Nonreactive particles are released at $\mathbf{X}_0 = (x_0, y_0) = (0.1, 0)$, which is located near the upstream boundary (see to Fig. 1). We are interested in the probability that a particle released at $\mathbf{X}_0$ reaches a particular region Ω_0 in the domain.

In this example, we take Ω_0 as a segment on the downstream boundary, $\Omega_0 = \{(x, y) | x = L_1, y \in [y_1, y_2]\}$, where L_1 is the domain size in the x direction. We assume that the velocity component in the longitudinal direction is deterministic, i.e., $c_1 = 0$, which allows us to derive an analytical solution as a benchmark for evaluating the computational efficiency of the CMC and ISMC methods. In fact, the spatial distribution of particles along the downstream boundary should follow a normal distribution and the probability of particle's reaching the region Ω_0 can be derived as

$$P(y \in \Omega_0) = \frac{1}{2}\left[erf\left(\frac{y_2}{\sqrt{2DL/v_1}} \right) - erf\left(\frac{y_1}{\sqrt{2DL/v_1}} \right) \right] \tag{34}$$

where $L = L_1 - x_0$ is the travel distance in x direction, erf the error function, and $D = c_2^2$ is velocity variance in y direction. In this example, we choose $D = 2.0$, i.e., the velocity fluctuation in the transverse direction, v_2, follows a normal distribution $N(0, 2)$.

Conventional Monte Carlo simulations are conducted for a large number of realizations. For each realization, a particle is released from point $\mathbf{X}_0$ and at each time step it is moved by a mean velocity of v_1 in the x direction and a velocity component $v_2(t)$ sampled from a normal distribution $N(0, 2)$. The particle's displacements in both directions are recorded until it hits the right boundary. In particular, the score function θ is recorded as one if the particle reaches Ω_0 or zero otherwise. This process is repeated for many realizations and the number of particles reaching the target region Ω_0 is used to approximate the probability $P = N_{\Omega_0}/N$, where N is the total number of simulations and N_{Ω_0} is the number of particles reaching Ω_0. Certainly, the number of simulations N should be large enough to achieve a statistically meaningful approximation of P. The estimated probabilities for $N = 1000$, 10000, and 100000, are illustrated in Figure 2. The analytical solution is also compared in the figure. Note that the curves from Monte Carlo simulations are shorter than that from the analytical solution because for regions that are far away from the central line, it is very difficult for particles to reach these regions and the estimated probability using Monte Carlo simulations is essentially zero. For example, if we take only 1000 samples (realizations), no particle has displacement larger than 135 on the downstream boundary and therefore the estimated probability is zero for all $y \geq 135$. As the number of simulations increases, some particles may have larger displacement in the y direction. And also, for a given distance from the central line, an increase of Monte Carlo simulations improves the accuracy of the estimation.

The procedure for conducting importance sampling Monte Carlo simulations is similar to that in the CMC as described above, except that an extra velocity component Δv is added to the original velocity field in the y direction. This means that the velocity in the y direction is $\Delta v + v_2(t)$, where $v_2(t)$ is taken from $N(0, 2)$ at each time step. The efficiency of the importance sampling method depends on the magnitude of this added velocity component Δv. In the case of $\Delta v = 0$, the importance sampling simulation is the same as the CMC method and the efficiency gain is one.

The optimized velocity component Δv may be determined from the release location, the target domain, and the mean velocity in the x direction. For example, if $\Omega_0 = \{(L_1, y)|y \in (200, 210]\}$, a segment of length 10 and centered at $y_0 = 205$, denoted as $L_{10}(205)$, on the right boundary, and if mean velocity $v_1 = 1.0$, we have $\Delta v = y_0 v_1/L_1 = 0.205$. Figure 3 compares particles' trajectories for the first 20 realizations in the CMC (dashed lines) and ISMC (solid lines) methods for Ω_0 given above. Without adding Δv, it is very hard for particles to reaching this region, because the probability is about 2.54×10^{-6}. If we need the estimation to be within 10% accuracy (relative error less than 10%), we have to conduct at least 3.93×10^7 Monte Carlo simulations. If we move Ω_0 further away from the central line ($y = 0$), for example, for $L_{10}(295)$, even more simulations are required.

By adding a velocity component Δv, particles move toward to the target region. In the example above, two (heavy lines) out of the first 20 realizations reach the target region. Of course, the probability is not one-tenth but needs to be corrected as detailed in the previous sections. Using the ISMC method with only $2,000$ simulations, we are able to derive a very accuracy estimation of 2.49×10^{-6}, i.e., a relative error of about 2% for

$L_{10}(205)$ for Δv given above. The advantage of the ISMC becomes even more obvious for much smaller probability.

Results from the ISMC method with 2,000 samples are compared in Figure 2 against those from analytical solutions and from the CMC method with different numbers of realizations. The figure shows that the ISMC method with 2,000 samples can accurately estimate the probability as low as in the order of 10^{-11}.

The estimation errors for the CMC can be computed from $\sigma_{MC}^2 = P(1-P)$, where P is the probability which can be computed from analytical solutions for this case. For the ISMC, the estimation variance is computed from (19). The comparison of the estimation variance for the CMC and ISMC is depicted in Figure 4. The figure shows that the estimation variance from the ISMC is substantially smaller than that from the CMC method.

Because the numbers of required simulations, N_{MC} and N_{ISMC}, for any given error ϵ are proportional to the estimation variance, Figure 4 indicates that the ISMC method is computationally much more efficient than the CMC method. These two numbers can be computed directly from $N_{MC} = \sigma_{CMC}^2/\epsilon^2$ and $N_{ISMC} = \sigma_{ISMC}^2/\epsilon^2$. The efficiency gain for the ISMC method is the ratio of these two numbers $\gamma = N_{MC}/N_{ISMC}$. Figure 5 compares the number of required samples as a function of distance from the central line for both the CMC and ISMC methods. While N_{CMC} increases quickly as the increase of the distance from the central line, N_{ISMC} is more or less a constant, ranging from 1,000 for $L_{10}(5)$ to 1,460 for $L_{10}(295)$. The figure also shows the efficiency gain of the ISMC method comparing to the CMC method. The gain ranges from about one for the region

436 $L_{10}(5)$ to as high as 10^9 for the region $L_{10}(295)$. The advantage of the ISMC method
437 becomes more obvious when the probability to be estimated is very small.

6. Conclusion and Discussion

438 This study is an extension of *Lu and Zhang* [2003a], in which the importance sampling
439 method was applied to simulate one-dimensional fluid flow and solute transport problems
440 with randomly constant properties. In this study we applied importance sampling Monte
441 Carlo simulations (ISMC) to estimating the probability of released conservative particles
442 reaching some particular regions of interest in multidimensional domains with correlated
443 medium properties. This problem may be solved by the conventional Monte Carlo simu-
444 lations (CMC), in which samples are taken from an ensemble of all possible trajectories
445 or paths (due to the heterogeneities of the porous media). Because the number of re-
446 quired simulations is inversely proportional to the probability to be estimated, the CMC
447 method is computationally expensive and in some cases it is nearly impossible to estimate
448 the probability if its value is very small. Although in many hydrologic applications, we
449 may not be very interested in the cases of very small probability, we do not know in
450 advance how small the probability is until the value is actually estimated. Therefore, it is
451 important to accurately estimate probability associated with events with low probability
452 values.

453 In the ISMC method, samples are taken from a biased distribution rather than the
454 original one such that more samples will be taken in the "important" regions of the
455 probability space. The results are then exactly corrected by a weighted average. In
456 particular, for transport of conservative solute particles, the ensemble of all trajectories
457 is modified by adding an extra velocity field to the original velocity field such that in the

new velocity field the particles will move toward the target regions. To preserve the mean, we need to make corrections by computing the weighted average.

The ISMC method has been demonstrated using two examples. In the first example, we revisited a one-dimensional transport problem presented in *Lu and Zhang* [2003a] for estimating the probability of a particle's travel time less than a given value. In the second example, particles were released in a two-dimensional domain with a random velocity field in the direction perpendicular to the mean flow direction. We chose these two simple examples because analytical solutions are available for both examples and they can sever as benchmarks for assessing the accuracy and efficiency of the CMC and ISMC methods.

The computational efficiency of the ISMC method is measured by the efficiency gain, which is the ratio of the numbers of required simulations in the CMC and ISMC methods. The efficiency of the importance sampling method depends on the choice of the biased distribution, which can be determined from the original mean flow field, and the locations of the release point and the target region. A different biased distribution may be needed for a different target region. Results from two demonstrative examples show that the ISMC method is computationally much more efficient than the CMC method and the efficiency gain can be many orders of magnitudes.

Acknowledgments. This work is partially supported by Los Alamos National Laboratory through a Laboratory Directed Research and Development (LDRD) project (20070441ER).

References

Chen J, D. Lu, J. S. Sadowsky, and K. Yao, On importance sampling in digital communications – Part I: Fundamentals, *IEEE J. Sel. Areas Commun.*, 11(3), 289-299, 1993.

Chen, M., A. A., Keller, D. Zhang, Z. Lu, and G. A. Zyvoloski, A stochastic analysis of transient two-phase flow in heterogeneous porous media, *Water Resour. Res.*, 42, W03425, doi:10.1029/2005WR004257, 2006.

Dagan G., Models of groundwater flow in statistically homogeneous porous formations, *Water Resur. Res.*, 15(1), 47-63, 1979.

Dagan G., Stochastic modeling of groundwater flow by unconditional and conditional probabilities, 1. Conditional simulation and the direct problem, *Water Resour. Res.* 18(4), 813-833, 1982.

Dellago, C., P. G. Bolhuis, F. S. Csajka, D. Chandler, Transition path sampling and the calculation of rate constants, *J. Chem. Phys.*, 108(5), 1964-1977, 1998.

Gelhar L. W., Stochastic Subsurface Hydrology. Prentice-Hall. Englewood Cliffs, NJ; 1993.

Glynn, P. W., and D. L. Iglehart, Importance sampling for stochastic simulations. *Manage. Sci.*, 35, 1367-1392, 1989.

Guadagnini, A., and S. P. Neuman, Nonlocal and localized analyses of conditional mean steady state flow in bounded, randomly nonuniform domains, 1, Theory and computational approach, *Water Resour. Res.*, 35(10), 2999-3018, 1999.

Indelman, P., and Y. Rubin, Solute transport in nonstationary velocity fields, *Water Resour. Res.*, 32, 1259-1267, 1996.

Li, S., and D. B. McLaughlin, A nonstationary spectral method for solving stochastic groundwater problems: unconditional analysis, *Water Resour. Res.*, 27(7), 1589-605, 1991.

Li, H., and D. Zhang, Probabilistic collocation method for flow in porous media: Comparisons with other stochastic methods, *Water Resour. Res.*, 43, W09409, doi:10.1029/2006WR005673, 2007.

Lu, Z., and D. Zhang, On importance sampling Monte Carlo approach to uncertainty analysis for flow and transport in porous media, *Adv. Water Resour.*, 26, 1177-1188, 2003a.

Lu, Z., and D. Zhang, Solute spreading in nonstationary flows in bounded heterogeneous saturated-unsaturated media, *Water Resour. Res.*, 39(3), 1049, doi:10.1029/2001WR000908, 2003b.

Lu, Z., and D. Zhang, Conditional simulations of flow in randomly heterogeneous porous media using a KL-based moment-equation approach, *Adv. Water Resour.*, 27(9), 859-874, 2004a.

Lu, Z., and D. Zhang, A comparative study on quantifying uncertainty of flow in heterogeneous media using Monte Carlo simulations, the conventional and the KL-based moment-equation approaches, *SIAM J. Sci. Comput.*, 26(2), 558-577, 2004b.

Lu, Z., and D. Zhang, Analytical solutions of statistical moments for transient flow in two-dimensional bounded, randomly heterogeneous media, *Water Resour. Res.*, 41, W01016, doi:10.1029/2004WR3389, 2005.

Lu, Z., and D. Zhang, Accurate, efficient quantification of uncertainty for flow in heterogeneous reservoirs using the KLME approach, *SPE Journal*, 11(2), 10.2118/93452-PA,

239-247, 2006.

Lu, Z., and D. Zhang, Stochastic simulations for flow in nonstationary randomly heterogeneous media using a Karhunen-Loeve based moment-equation approach, *SIAM Multiscale Model. Simul.*, 6(1), 228-245, 2007

Mazonka, O., C. Jarzynski, and J. Bloki, Computing probabilities of very rare events for Langevin processes: a new method based on importance sampling, *Nucl. Phys. A*, 641, 335-354, 1998.

Osnes, H. Stochastic analysis of head spatial variability in bounded rectangular heterogeneous aquifers, *Water Resour. Res.*, 31, 2981- 2990, 1995.

Sadowsky, J. S., On the optimality and stability of exponential twisting in Monte Carlo estimation, *IEEE Trans. Inform. Theory*, 39, 119-128, 1993.

Smith, P. J., M. Shafi, and H. Gao, Quick simulation: a review of importance sampling techniques in communications systems, *IEEE J. Sel. Areas Commun.*, 15(4), 597-613, 1997.

Sun, A. Y., and D. Zhang, Prediction of solute spreading during vertical infiltration in unsaturated, bounded heterogeneous media, *Water Resour. Res.*, 36, 715-723, 2000.

Tartakovsky, D. M., S. P. Neuman, and Z. Lu, Conditional stochastic averaging of steady state unsaturated flow by means of Kirchhoff transformation, *Water Resour. Res.*, 35(3), 731-745, 1999

Yeh, T.-C., L. W. Gelhar, A. L. Gutjahr, Stochastic analysis of unsaturated flow in heterogeneous soils: 1. Statistically isotropic media, *Water Resour. Res.*, 21(4), 447-456, 1985.

Zhang, D., and Lu, Z., An efficient, higher-order perturbation approach for flow in randomly heterogeneous porous media via Karhunen-Loeve decomposition, *J. Comput. Phys.*, 194(2), 773-794, 2004

Zhang, D., Numerical solutions to statistical moment equations of groundwater flow in nonstationary, bounded heterogeneous media, *Water Resour. Res.*, 34, 529-538, 1998.

Zhang, D., *Stochastic Methods for Flow in Porous Media: Coping with Uncertainties*. Academic Press, San Diego, Calif., 2002.

Table 1. Number of simulations required for different values of T_0

T_0	P_T	N_{MC}	P_{MC}	Δv	P_{IS}
300.	0.500	1.00×10^{2}	0.499 (1,000)	0.0	0.499 (1,000)
250.	4.78×10^{-2}	1.99×10^{3}	4.50×10^{-2} (2,000)	1/15	4.88×10^{-2}(1,000)
200.	4.29×10^{-4}	2.33×10^{5}	$3.78\times10^{-4}(233,000)$	1/6	4.40×10^{-4} (1,000)
150.	2.87×10^{-7}	3.49×10^{8}	-	1/3	3.06×10^{-7}(1,000)
100.	1.31×10^{-11}	7.64×10^{12}	-	2/3	$1.37\times10^{-11}(1,000)$
50.	3.93×10^{-17}	2.54×10^{18}	-	5/3	$3.46\times10^{-17}(1,000)$

Figure Captions

Figure 1 Schematic diagram illustrating the importance sampling method: estimating the probability of a solute particle released from a source reaching a particular region of interest, Ω_0.

Figure 2 Comparison of probability estimated from analytical solutions, the ISMC method with 2,000 samples, and the CMC method with different number of samples.

Figure 3 Twenty trajectories from the CMC method and their corresponding trajectories from the ISMC method for estimating the probability that particles reaches the region $L_{10}(205)$.

Figure 4 Comparison of estimation variance of the CMC and ISMC methods.

Figure 5 The number of required samples for the CMC and ISMC methods and the efficiency gain of the ISMC method as functions of the distance from the central line.

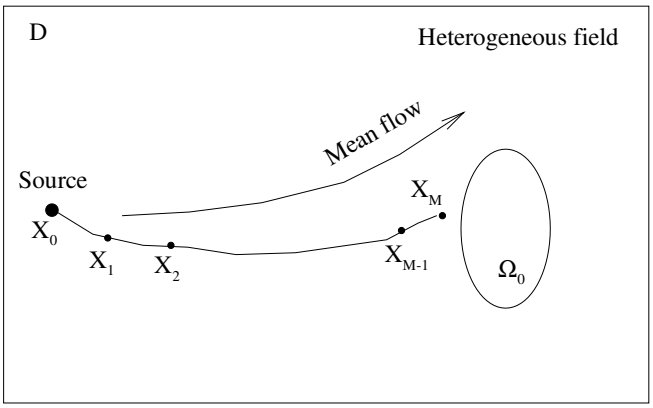

Figure 1. Schematic diagram illustrating the importance sampling method: estimating the probability of a solute particle released from a source reaching a particular region of interest, Ω_0.

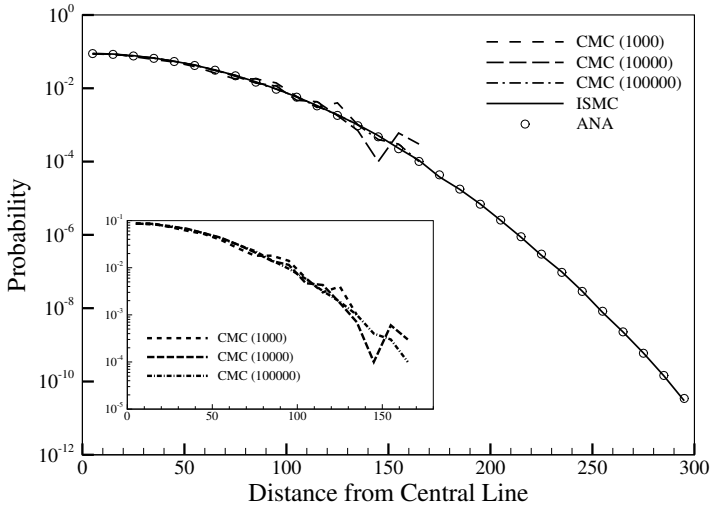

Figure 2. Comparison of probability estimated from analytical solutions, the ISMC method with 2,000 samples, and the CMC method with different number of samples.

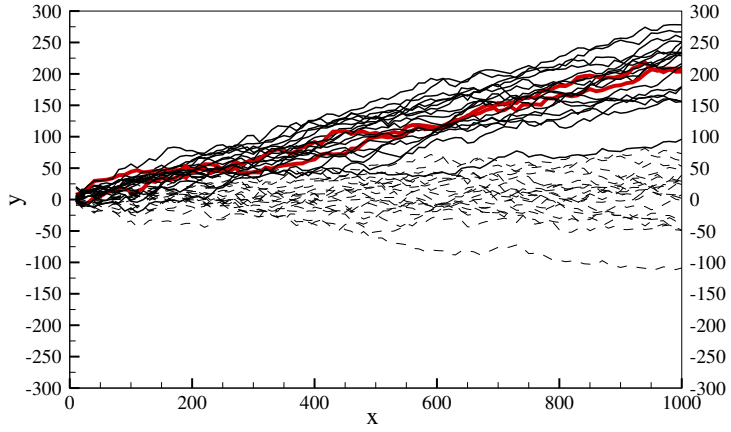

Figure 3. Twenty trajectories from the CMC method and their corresponding trajectories from the ISMC method for estimating the probability that particles reaches the region $L_{10}(205)$.

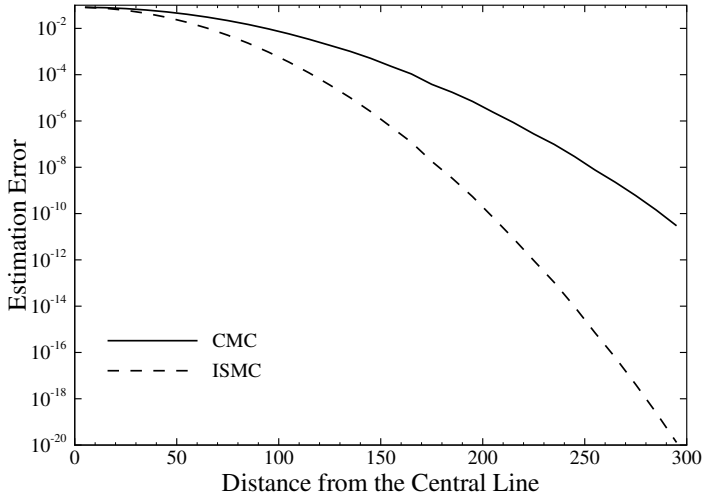

Figure 4. Comparison of estimation variance of the CMC and ISMC methods.

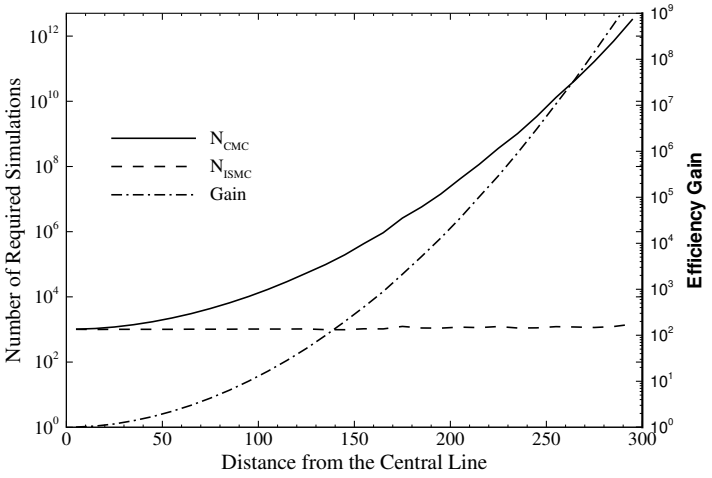

Figure 5. The number of required samples for the CMC and ISMC methods and the efficiency gain of the ISMC method as functions of the distance from the central line.